Annotation Imputation to Individualize Predictions: Initial Studies on Distribution Dynamics and Model Predictions

London Lowmanstone[1,*,†], Ruyuan Wan[2,†], Risako Owan[1], Jaehyung Kim[3] and Dongyeop Kang[1]

[1]University of Minnesota
[2]University of Notre Dame
[3]KAIST

Abstract

Annotating data via crowdsourcing is time-consuming and expensive. Due to these costs, dataset creators often have each annotator label only a small subset of the data. This leads to sparse datasets with examples that are marked by few annotators. The downside of this process is that if an annotator doesn't get to label a particular example, their perspective on it is missed. This is especially concerning for subjective NLP datasets where there is no single correct label: people may have different valid opinions. Thus, we propose using imputation methods to generate the opinions of all annotators for all examples, creating a dataset that does not leave out any annotator's view. We then train and prompt models, using data from the imputed dataset, to make predictions about the distribution of responses and individual annotations.

In our analysis of the results, we found that the choice of imputation method significantly impacts soft label changes and distribution. While the imputation introduces noise in the prediction of the original dataset, it has shown potential in enhancing shots for prompts, particularly for low-response-rate annotators. We have made all of our code and data publicly available.[1]

Keywords

natural language processing, imputation, matrix factorization, content filtering, large language models, annotation, NLPerspectives, LeWiDi

1. Introduction

Natural language processing (NLP) models rely on large amounts of data that is expensive and time-consuming to label [1]. Crowdsourcing has emerged as a popular solution to this problem, but it comes with its own challenges, principal among them being annotator disagreement [2, 3]. Although there are many possible causes of disagreement, the common causes are annotator subjective judgment and language ambiguity [4]. Not taking into account the inherent subjectiveness and ambiguity of some instances can lead to inaccurate predictions [5]. Thus, in recent years, researchers have begun to recognize the importance of disagreement, advancing models and datasets that accurately reflect disagreement, rather than ignoring it or working around it [6].

In order for models to accurately reflect disagreement, they must accurately model true human populations. Here, we frame the problem of making accurate predictions for individual annotators as an *imputation* problem: given a spreadsheet with rows corresponding to text and columns corresponding to annotators, how would one accurately fill in the spreadsheet in order to correctly predict how each annotator will label each piece of text? Figure 1 visualizes this approach, which, ideally, enables dataset creators to generate additional annotations without extensive crowdsourcing.

We postulate that annotators who have historically assigned the same labels to identical text segments may, given similar contexts in unseen data, continue to demonstrate congruent labeling behavior. Thus, imputation methods, which take in data containing all of the dataset annotations, should be able to discover patterns to relate annotators and annotations in order to make accurate predictions as to how a particular annotator might label a particular example, based on how other annotators labeled the same or similar examples.

Matrix factorization techniques used in recommendation systems and annotator-level models of disagreement both make predictions about individual annotations made by individual annotators. Thus, our analyses can be applied to both types of models in order to reveal differences between the original data and imputed data created by these models. In our work, we impute datasets by utilizing two matrix factorization methods, kernel matrix factorization and neural collaborative filtering, and a supervised learning model (Multitask) proposed by [7], that

2nd Workshop on Perspectivist Approaches to NLP

*Corresponding author.
†These authors contributed equally.
lowma016@umn.edu (L. Lowmanstone); rwan@nd.edu (R. Wan); owan002@umn.edu (R. Owan); jaehyungkim@kaist.ac.kr (J. Kim); dongyeop@umn.edu (D. Kang)
0000-0002-5457-6553 (L. Lowmanstone)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://github.com/minnesotanlp/annotation-imputation

CEUR-WS.org/Vol-3494/paper10.pdf

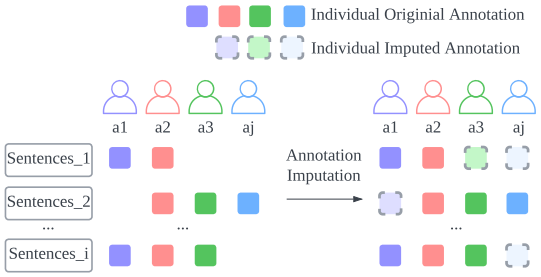

Figure 1: Annotation imputation by using individualized prediction. Each square represents a single annotation. The original dataset on the left is missing some annotations from annotators. We then make predictions as to how each of the missing annotations would be filled in, resulting in the imputed dataset on the right. The slightly transparent squares indicate imputed annotations that are not in the original dataset. We then analyze how the imputed dataset on the right differs from the original data on the left.

models disagreement at the annotations level [8, 9, 7]. Through our analyses, we find that imputation greatly transforms the distribution of annotations (including lowering the variance of the data) and creates noticeable changes in examples' soft labels.

After imputing and analyzing the data, we use the imputed datasets to train and prompt models that make individualized predictions. For training, we use the Multitask model from [7] in order to make aggregate and individualized predictions and find that training on imputed data harms prediction performance. For prompting, we use GPT-3 (text-davinci-003) and ChatGPT (3.5-turbo) and provide the models with prompts containing either imputed or non-imputed data to determine their impact on the models' ability to make individualized and distributional label predictions. We find that adding prompt shots via imputation improves ChatGPT's performance for predicting annotations of low-response-rate annotators, but does not consistently improve other areas of prediction such as distributional label prediction, individualized prediction for high-response annotators, or merely replacing human annotations with imputed data [10].

In summary, our primary contributions are:

1. Framing individualized prediction as an imputation problem
2. Analysis techniques to compare imputed data to real data:
   a) *Distribution Analysis*, which focuses on transformations of the underlying distribution of annotations after imputation. We show that different imputation methods significantly change the underlying annotation distributions.
   b) *Soft Label Analysis*, which focuses on shifts in the soft label after imputation compared to the original data. We provide a visualization technique for viewing how the soft labels change after imputation.
   c) *Usage Analysis*, which focuses on how models perform after training on or being prompted with imputed data. We show that kernel matrix factorization, neural collaborative filtering, and Multitask imputation tend to harm the capabilities of Multitask and GPT models to make individual, soft-label, and aggregate predictions, except in the case of using imputation to increase the number of shots to prompts for making individualized predictions for low-response-rate annotators.

2. Related Work

Disagreement in NLP Disagreement has been found within NLP datasets for many years [11, 12, 13, 6]. However, recently, there has been much work done on developing and evaluating models that model disagreement within datasets, rather than ignore disagreement [7, 14, 5, 15, 16].

In particular, the SemEval-2023 Learning with Disagreements (LeWiDi) task invites competitors to create models that predict soft labels of human disagreement for different text inputs [6]. While hard labels provide a definitive categorization for data points, soft labels offer a probabilistic interpretation, capturing the uncertainties or nuances in classification. Multiple submissions for this task used models proposed by [7] in order to make predictions at the individual level. Success at the task was determined by micro F1 score on gold labels and cross-entropy on soft-labels. Within the task, all dataset labels were binary, and no metric was used to measure success at the level of individual annotators.

The authors of [17] propose multiple different methods for evaluating models that make individualized predictions. Among these are Jensen-Shannon divergence, a symmetric variation of Kullback-Leibler (KL) divergence and cross-entropy. F1 score is also a proposed metric, but only for aggregate labels, not individual labels.

Another model for approaching disagreement is Jury Learning, where individuals' annotations are modeled in order to form "juries" of different demographics [14]. In their paper, the authors analyze how using data generated by "juries" affects the aggregate label, particularly in the case of contentious texts [14].

Collaborative Filtering in Recommendation Systems Similar to modeling disagreement, collaborative

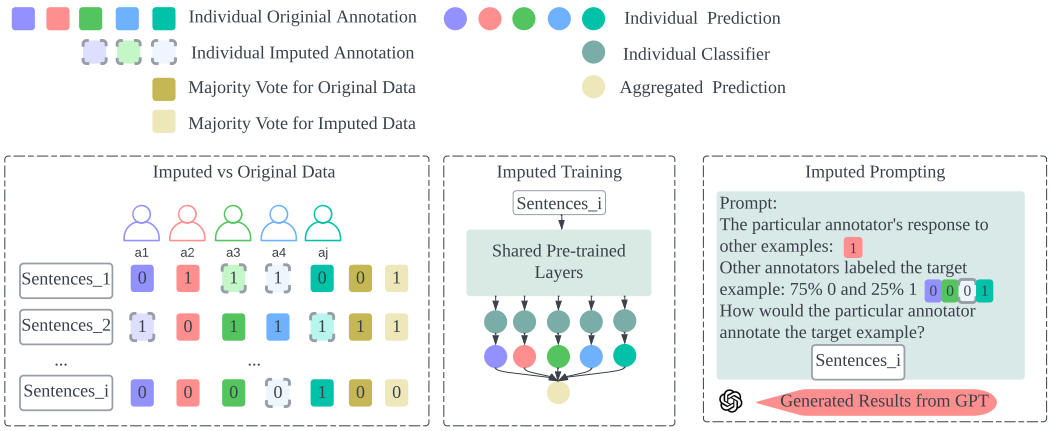

Figure 2: Three experiments of annotation imputation: (1) comparing imputed vs original data, (2) training on imputed data (3) generating with prompts based on imputed data.

filtering systems also create individualized predictions of human behavior in order to make relevant recommendations. Contrary to disagreement models in natural language processing, these systems are entirely dependent on user-provided annotations and lack the ability to predict the reactions of new users to unseen text.

When evaluating performance of collaborative filtering systems, metrics are generally focused on accuracy of predictions, rather than quantifying and visualizing changes in the distribution of data [18, 19]. These metrics provide good signals for the success of a model, but do not help with understanding how models modify data when they do not match the original data.

3. Annotation Imputation For Individualized Predictions

First, we compared how various imputation methods handle and fill in the missing annotations. We then trained supervised models and used GPT-based prompting to evaluate imputation's impact on aggregate and individualized prediction.

3.1. Annotation Imputation

In order to understand how individualized prediction affects data, we use three different methods: kernel matrix factorization, neural collaborative filtering, and a Multitask supervised neural network model from [7].[1]

[1]The hyperparameters used for each of the models can be found in Appendix C.

Kernel matrix factorization relies on kernels to project data to a higher dimensional space where more complex patterns can be found in order to generate a matrix factorization which is used for imputation. NCF matrix factorization relies on neural networks rather than kernels to compute a matrix factorization of the data, and Multitask relies purely on neural networks to make individualized predictions. All methods employ a core process: identifying patterns between annotators and annotations across the dataset.

For our experiments, kernel matrix factorization is implemented primarily using off-the-shelf code [20]. In addition, we add a grid search component which determines the best model hyperparameters by holding out 5% of the given training data as validation data, and choosing the hyperparameters that resulted in the lowest RMSE score on the validation data. See Appendix C for details.

Neural collaborative filtering was implemented based on the work of [9]. The details of our implementation can be found via our code. For this model, we also use an additional grid search component which determines the best model hyperparameters. However, we choose the hyperparameters for this model based on the lowest RMSE score when evaluated on all training examples, rather than a held-out validation set. See Appendix C for details.

3.2. Imputed Training

In this stage, we use the Multitask model from [7] on both original and imputed data and compare the evaluation results in order to understand how imputed data impacts model training. We follow a similar setup to

[7] by using 5-fold validation and averaging the results across the folds [7]. However, in order to account for dataset imbalance in our datasets, we report weighted F1 scores, rather than macro F1 scores. Note that the data from each validation fold is hidden from the imputer, so as not to cause data leakage. Details of the model's architecture can be found in Appendix A, and hyperparameter details can be found in Appendix C. The same model is used both for imputation and training (see Section 4).

3.3. Imputed Prompting

We also conducted three key experiments using GPT-3 (text-davinci-003) and ChatGPT (3.5-turbo) to better understand the impact of imputation on predictions made by GPT-based models [10]: The first experiment tests the impact of using imputed data when making individualized predictions for low-response-rate annotators. The second experiment makes individualized predictions for all annotators (not just low-response-rate annotators), but also adds original distribution information, imputed distribution information, or the original majority-voted label near the end of the prompt in addition to the included individual examples to quantify the impact of the extra information on predictions. The third tests individualized predictions when either original or imputed data from three distinct annotators is provided in the prompt. Of these, imputation only had a positive impact on making individualized predictions for low-response-rate annotators; the other two experiments are included in Appendix D.

For all experiments, we create prompt skeletons, which are then filled in with data and/or text, depending on the experiment run (see Appendix F). This enables us to understand the influence of different prompts and data.

Individualized Predictions for Low-Response-Rate Annotators In this experiment, we first isolated from each dataset the 30 annotators with the lowest number of annotations in the dataset. We then generated a prompt for each of those 30 annotators. Each prompt consists of at most 30 sentences and annotations from that annotator (if there were more, we discarded the extras and chose one to hold out, and if there were less, we included all but one to hold out). Following the real examples, we also included an additional 30 examples whose sentences are from the dataset (and differ from the previous 30 examples and the held-out example), but whose annotations are imputed via NCF. The final section of the prompt then asks ChatGPT to predict the annotator's annotation on the held-out example.

In the experiment, we test for differences between three different conditions:

1. Including both the original and imputed data
2. Only including the original data
3. Only including the imputed data

In each of these conditions, outputs are considered correct if, after removing whitespace, they only contain the correct label. We conducted initial studies to discard particularly low-performing skeletons and infills. The remaining skeletons and infills are used for all conditions. (Details are provided in our code.) We then measure success of a condition based on the highest weighted F1 score achieved by a prompt skeleton within that condition.

4. Experiments

Our experiments involve: (1) comparing imputed and original data, (2) conducting training using imputed data, and (3) prompting generation based on imputed data, all illustrated in Figure 2.

Datasets In order to ensure a diversity of data, we utilize six different datasets in our analysis: Social Chemistry (SChem) [21], Social Bias Inference Corpus (SBIC) [22], Gab Hatespeech Corpus (GHC) [23, 24], Sentiment dataset [25], and Politeness dataset [26]. Additionally, we isolate examples from the SChem dataset that were labelled by 5 annotators in order to form the SChem5Labels dataset. Our datasets are summarized in Table 1, and more details can be found in Appendix B.

4.1. Imputed vs Original Data

Imputation We impute each of the datasets with each of the imputation methods. However, in order to judge which methods have the best performance, we also test imputing the data while withholding 5% of the annotations for evaluation. Withheld data is chosen in a manner that reduces duplicate examples and annotators within the withheld data in order to provide a more diverse test set (details can be found in our code).

Table 2 summarizes the RMSE score for each of the methods on each of the datasets when evaluated on the withheld data. Note that the Politeness dataset collects labels ranging from 1 to 25, implying a broader variance compared to other datasets. Consequently, RMSE values are expected to be higher for the Politeness dataset. We also find that while Multitask and NCF perform best on different datasets, kernel matrix factorization is never the best method, and is in fact always dominated by the NCF method.

After the data is imputed, we use two analyses in order to better understand how imputed data differs from original data.

Dataset	# instances	# annotators	# annotation	Label Types
SChem	400	100	50	No one believes (0), occasionally believed (1),
SChem5Labels	8007	102	5	controversial (2), common belief (3), universally true (4)
SBIC	45223	304	3	Not offensive (0), maybe (0.5), offensive (1)
GHC	27538	18	3-4	Not hate speech (0), hate speech (1)
Sentiment	14070	1481	4-5	Very negative (-2), somewhat negative (-1), neutral (0), somewhat positive (1), very positive (2)
Politeness	4338	219	5	A scale from polite (1) to impolite (25).

Table 1
Statistics and label information on the six datasets we use across our analyses. The statistics include the number of unique text instances, the number of unique annotators, and the number of annotations per text instance in the six datasets.

Method	SChem	SChem5Ls	GHC	SBIC	Sentiment	Politeness
Multitask	0.82	0.72	**0.32**	**0.64**	1.14	4.41
NCF	**0.63**	**0.66**	0.35	0.65	**0.90**	**3.69**
Kernel	0.71	0.72	0.36	0.90	1.03	4.39

Table 2
RMSE scores of the different imputation methods across datasets. All models were run once except kernel matrix factorization, whose reported scores are the median of 3 runs with differing random seeds. The lowest RMSE score on each dataset is in **bold**, and the second-lowest is underlined.

Distributional Analysis The first analysis (distribution analysis) applies principal component analysis (PCA) to both imputed and original data to visualize shifts in the distribution of example ratings. In order to apply PCA, we represent each text as a vector of its annotations, where missing annotations are filled in with a value of 10, which is far outside the range of valid annotation labels for these datasets [27]. We also calculate the change in variance between imputed and original data, and graph this variance against the disagreement rate across examples. The disagreement rate is computed as the number of annotations that disagree with the majority-voted label for that example, divided by the total number of annotations for that example. The majority-voted label for imputed data is computed on the imputed data.

When we project the annotations to two dimensions using PCA, we find that different imputation methods cause significant changes to the distribution of the data as shown in Figure 3.[2] Each imputation method generates an extremely different underlying distribution for the annotations.

In addition, we compute how the variance and disagreement rate change after imputation with NCF matrix factorization. Our results are compiled in Table 3, and we also provide Figure 4 to display the results on the SChem dataset. Results from other methods can be found in Appendix H. Across all datasets, we find that imputation decreases variance, indicating that NCF matrix factorization does not accurately model the diversity of human annotations. We can observe this lowered variance in both Figure 3 and Figure 4 by comparing the scale of the plots in the PCA visualization and by comparing the heights of the points in the variance plot. We also find

[2]Other datasets' results can be found in Appendix G.

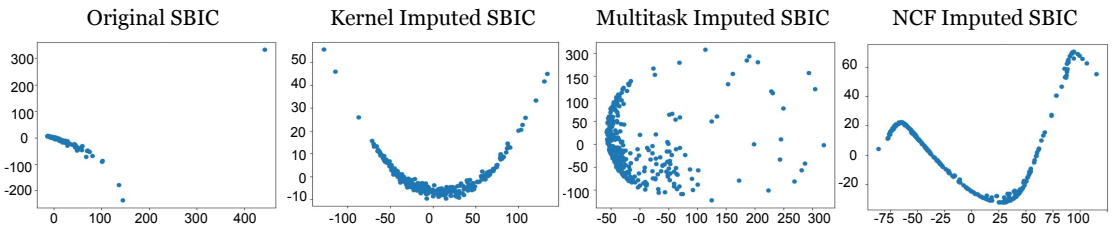

Figure 3: PCA projection of the SBIC dataset before and after using imputation.

Statistic	SChem	SChem5Labels	Politeness	Sentiment	SBIC	GHC
Avg Change: Variance	**-0.096**	**-0.088**	**-6.987**	**-0.312**	**-0.044**	**-0.004**
Avg Change: Disagreement Rate	**-0.061**	**-0.060**	0.18	**-0.106**	0.044	**-0.006**

Table 3
Change in average variance and disagreement rate due to using NCF matrix factorization to impute the dataset. Instances where the variance or disagreement rate are **lowered** due to imputation appear in bold.

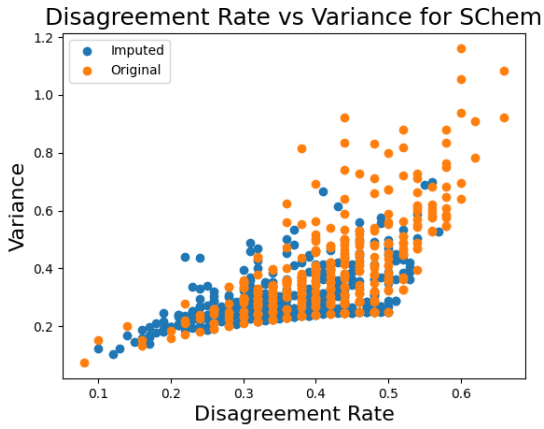

Figure 4: A graph displaying how the variance has decreased after using NCF matrix factorization. Each point represents an example. Variation is across annotations for that example, and disagreement rate is the percentage of people who disagree with the majority-voted annotation.

that NCF matrix factorization tends to, but does not always, lead to more agreement with the majority-voted annotations.

Overall, the chosen method for individualized prediction has a large impact on the structure underlying the predictions, even within the same dataset. We also find imputers can lower variance and raise agreement within the dataset, demonstrating that imputation models may not always capture the diversity and disagreement of real human annotators.

Soft Label Analysis The second analysis visualizes differences between soft labels of examples between the original and imputed data. To create the visualization, we assign each label to a color and then generate horizontal bars of equal size where the proportion of the bar containing that color corresponds to the proportion of annotations with that label. This enables us to directly compare how different imputation methods alter the soft label distribution. Similar to [17], we also calculate the Kullback-Liebler (KL) divergence between the original distribution of data and the imputed data in order to numerically quantify the difference between distributions.

Through our soft label analysis, we find that different imputation methods lead to varying changes of the soft label of examples after imputation. Figure 5 demonstrates how imputation changes the distribution of the data for a given example and allows one to directly compare different imputation methods to see how they modify the data. In the case of Figure 5, we see an example from the SChem dataset which shows that kernel matrix factorization predicted a much smaller proportion of annotators to give the highest rating (in pink) than was in the original dataset, while NCF matrix factorization predicted a moderately larger proportion of users to give the second-highest rating (in blue) than the original dataset.

Since we are interested in understanding how these soft labels differ from the original data, we also compute the KL divergence between the imputed data and the original data. Note that if one would like a symmetric metric, the Jensen-Shannon divergence could be computed here as well [17]. In the particular example in Figure 5, we see that the KL divergence score on Example 97 for Kernel is 0.105, compared to the 0.123 divergence score by NCF. We provide a selection of multiple examples in Appendix I. We also provide the average and standard deviation of KL divergence from the original data for each dataset and each imputation method in Table 4. Overall, we see that NCF matrix factorization tends to best preserve the soft label of the original dataset when compared to kernel matrix factorization and the multitask model, as it is always either best or second-best. However, performance is dataset-dependent, and kernel and multitask achieve the best fidelity to the original soft labels for the Sentiment and SBIC datasets, respectively.

Overall, soft labels do not remain consistent through imputation, and some methods of individualized prediction may tend to better preserve soft labels than others. In our case, NCF matrix factorization best preserved the soft labels.

4.2. Imputed Training

For imputed training, we train the Multitask model on the original data, data imputed by NCF, and data imputed by a separate Multitask model. (Since RMSE scores from kernel matrix factorization are worse than NCF on each

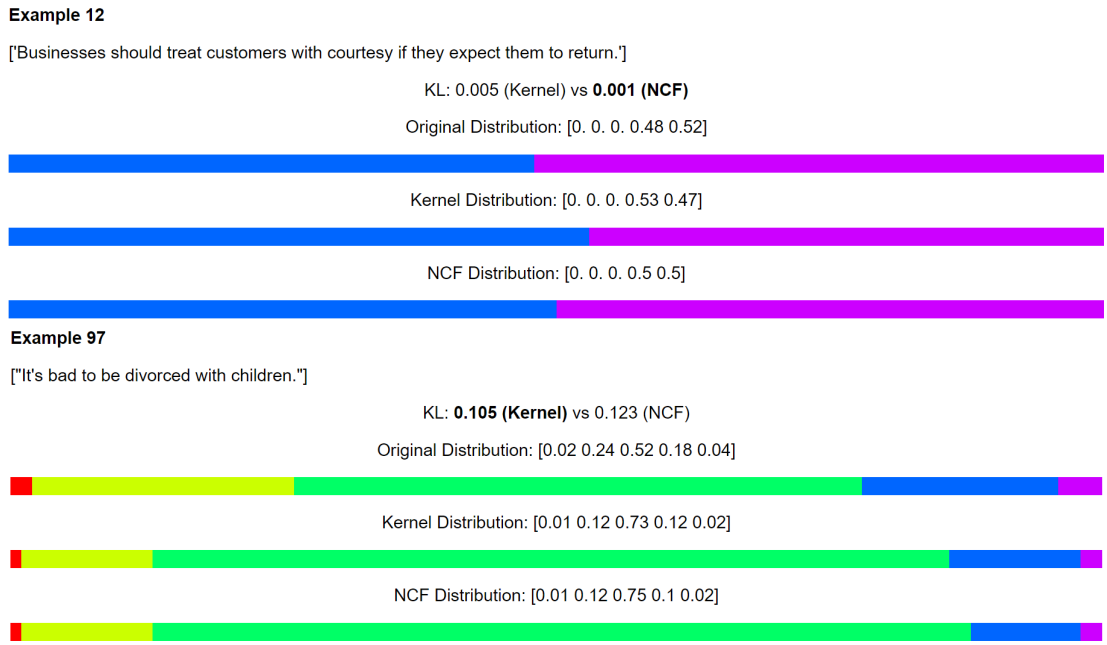

Figure 5: Visualizations showing the shift in distribution between the original distribution (soft label), the distribution after using kernel matrix factorization, and the distribution after using NCF matrix factorization on examples within the SChem dataset. The exact proportions of each label are listed, as well as the KL divergence score for each method at the top, with the method best reflecting the original distribution in **bold**. The top Example 12 illustrates an example for which both NCF and kernel matrix factorization do a good job of keeping the original soft-label, whereas for the bottom Example 97, both methods highly over-estimate the proportion of annotators who respond with the middle label.

Method / Dataset	SChem	SChem5Ls	Sentiment	SBIC
NCF	**0.030**$\pm$**0.0262**	**0.233**$\pm$**0.305**	0.582$\pm$0.678	0.413$\pm$0.615
Kernel	0.036$\pm$0.0268	0.292$\pm$0.363	**0.709**$\pm$**0.700**	0.465$\pm$0.654
Multitask	0.080$\pm$0.109	0.317$\pm$0.369	0.540$\pm$0.359	**0.307**$\pm$**0.359**

Table 4
Average and standard deviation of the KL divergence across datasets and individualized prediction methods. The method which best preserved the original distribution is in **bold**, and the KL divergence of the second-best method is underlined. Note that preserving the soft label / distribution is not necessarily indicative of accuracy or performance.

dataset, we omit kernel matrix factorization from this experiment.) After training the Multitask model on the original and imputed data, we report the average and standard deviation of the weighted F1 score for individual and aggregate predictions over 5 folds using 5-fold validation in Table 5 and Table 6. We observe that training on imputed data from NCF and Multitask results in performance worse than if we had used just the original data. This indicates that the predictions made by each of the methods biases the data in a way that does not match the true predictions that the annotators would have made.

However, not all prediction models had the same level of performance, and different datasets observed different results. Generally, using the original data resulted in the best outcomes, followed by using the Multitask model to impute the training data. Using NCF matrix factorization to impute the data resulted in the lowest performance.

When we break out the model's performance to examine success on examples with differing levels of disagreement (Table 7), we see that the model tends to perform much better on examples with higher agreement among annotators. We also see that the drop in performance from imputing data is fairly consistent across disagreement levels, except for the GHC dataset anomaly on low disagreement examples, where imputation helped

Method / Dataset	Politeness	GHC	SChem
Original	**0.33**$\pm$**0.004**	**.89**$\pm$**0.003**	**0.53**$\pm$**0.009**
NCF	0.18$\pm$0.005	0.88$\pm$0.013	0.52$\pm$0.014
Multitask	0.31$\pm$0.007	**0.89**$\pm$**0.009**	0.52$\pm$0.006

Table 5
Average weighted F1 score of *individualized* predictions made by the Multitask classifier trained on data generated by either NCF matrix factorization or a separate Multitask model. All the values and error bars are mean and standard deviation across five folds. The best and the second best results on each dataset are indicated in **bold** and underline, respectively.

Method / Dataset	Politeness	GHC	SChem
Original	**0.38**$\pm$**0.019**	**.919**$\pm$**0.004**	**0.611**$\pm$**0.91**
NCF	0.22$\pm$0.007	0.912$\pm$0.010	**0.611**$\pm$**0.91**
Multitask	0.34$\pm$0.016	0.915$\pm$0.009	**0.611**$\pm$**0.91**

Table 6
Average weighted F1 score of *aggregate* (majority-voted) predictions made by the Multitask classifier trained on data generated by either NCF matrix factorization or a separate Multitask model. All the values and error bars are mean and standard deviation across five folds. The best and the second best results on each dataset are indicated in **bold** and underline, respectively.

slightly. How disagreement levels are computed is discussed in detail in Appendix J.

Overall, this indicates that different methods of individualized predictions can introduce different biases into the data that cause methods trained on these predictions to perform worse than if they had trained on just the original data. Since we expect performance to increase with the amount of data provided, we conclude that these particular methods of individualized prediction likely introduce strong biases that do not reflect reality [28].

4.3. Imputed Prompting

Here, we highlight the results of using imputed data to improve individualized predictions on low-response-rate annotators, as shown in Table 8. (As mentioned above, other experiments are detailed in Appendix D) From the data, we observe that using solely imputed data outperforms using original data or adding original data to the imputed data for all datasets except for Politeness.

Politeness is likely an outlier due to the high range of potential labels in the Politeness dataset, leading the NCF method to impute labels that are unlikely to occur in the real dataset, causing ChatGPT to also predict unlikely labels. However, when the amount of labels is smaller (5 or less), using only imputed data increases performance.

We conjecture that since low-response-annotators in these datasets generally have far less than 30 original annotations, imputation enables us to provide more shots to ChatGPT than the original dataset could provide, thus enabling more accurate predictions than can be made without imputation. While more data is needed to determine why combining both imputed and original data performs poorly, we provide supporting experiments in Appendix E to demonstrate that the performance improvement from using imputed data is particular to low-response-rate annotators and is caused by the imputed data, not the prompt text.

5. Discussion

Our analyses shed light on the impact of various imputation methods on the structure, soft label, and training/prompting viability of imputed data in the context of NLP annotation tasks in comparison to purely human-labeled data. We demonstrate that different imputation methods can lead to significantly different underlying distributions of the data, which can, in turn, affect the performance of models trained on this data. Furthermore, while imputation can introduce noise, diminishing the accuracy of predictions for the original dataset, it is essential to consider that the original dataset may not wholly capture the full spectrum of reality due to the absence of some annotator opinions. This has important implications for the design and evaluation of individualized prediction models in various applications, as well as for understanding and quantifying the biases that may be introduced by such models.

Each one of our analyses focuses on a particular area of interest, which, together, help researchers and practitioners to better understand the predictions made by individualized prediction models. The distribution analysis provides information to those who are interested in ensuring that their model's predictions match the distribution of the original data and tools for analyzing changes in disagreement and variation. For those who are interested in soft labels, such as competitors in future LeWiDi tasks, our visualization helps with understanding how models estimate the soft label and computational tools for determining which models mimic the original soft label best. As we see a rise in human-level predictions from systems, it is important to understand if models can be trained or prompted with data created by individualized prediction models. We provide analyses from base systems indicating how the chosen imputation method may affect performance. Regardless of the scenario, our provided analyses enable researchers and practitioners who use models that make individualized predictions to better understand the differences between their model's

Dataset	Not Imputed			Imputed		
	Disagreement	N	Value	Disagreement	N	Value
Politeness	Low	1306	**0.481**$\pm$**0.015**	Low	1306	0.352$\pm$0.024
	Medium	2267	0.293$\pm$0.013	Medium	2267	0.203$\pm$0.008
	High	762	0.193$\pm$0.013	High	762	0.121$\pm$0.005
GHC	Low	20344	0.965$\pm$0.004	Low	20344	**0.968**$\pm$**0.003**
	Medium	814	0.721$\pm$0.012	Medium	814	0.654$\pm$0.027
	High	6392	0.717$\pm$0.006	High	6392	0.654$\pm$0.021
SChem	Low	133	**0.608**$\pm$**0.043**	Low	133	0.604$\pm$0.050
	Medium	137	0.590$\pm$0.033	Medium	137	0.581$\pm$0.030
	High	130	0.404$\pm$0.050	High	130	0.394$\pm$0.045

Table 7
F1 values from individualized prediction done by the Multitask model, broken out by disagreement in the original dataset. The highest F1 score for each dataset is in **bold**, and the second highest is underlined. The "N" column signifies how many examples are in each category.

Method / Dataset	Politeness	GHC	SChem	SChem5L	SBIC	Sentiment
Combined	0.13	0.75	0.50	0.60	**0.95**	0.58
Original	**0.14**	0.85	0.53	0.49	0.93	0.31
Imputed	0.07	**0.86**	**0.56**	**0.65**	**0.95**	**0.60**

Table 8
Highest Weighted F1 score for predicting the annotations of 30 users with the lowest response rate in the dataset across multiple prompt skeletons and infills of those skeletons. Imputation is done via the NCF method. The best result for a dataset is in **bold**, while the second-best is underlined.

predictions and real human annotations.

6. Future Work

While we include two different matrix factorization methods from collaborative filtering, content-based recommendation systems also provide individualized predictions, so future work includes applying our methods to a content-based recommendation system.

Also note that each of the methods we use is not state-of-the-art in their respective field. We have chosen baseline models for ease of implementation. Future work includes running our methods on more advanced systems that may make more accurate predictions.

We also have not conducted a user study to verify and quantify that our analysis methods help with understanding how predicted data differs from original data. Our analysis here is based on the fact that previous methods rely on aggregate metrics and do not provide fine-grained and comparative data between original and predicted data. Conducting a user study would allow us to provide explicit evidence of the exact amount of improvement our methods provide in general for understanding how individualized prediction impacts data.

7. Limitations

While our methods are extendable to any model that makes individualized predictions, we only test our methods on baseline models for both disagreement modeling and collaborative filtering. Thus, when used on state-of-the-art methods, our methods may give very different results. However, we still expect these methods to be useful for understanding how imputation modifies the underlying data, even if those modifications do not match our results.

8. Conclusions

We have proposed and utilized four different methods of understanding how the predictions made by individualized prediction models differ from the original data. We found that for kernel matrix factorization, and NCF matrix factorization, the original soft label for the data shifts in different ways based on the method used, the variance in labels is overall lowered, and training on data created by these methods results in generally worse prediction performance, while imputed data can be used to increase the number of shots in prompts.

Overall, we hope that our analysis methods for models that make individualized predictions are applied to future models in order to help researchers and practitioners to

better understand how their models' predictions differ from real human annotators.

Ethics Statement

Any methods which attempt to make individualized predictions carry the risk of learning how to replicate aspects of individuals' identities in order to make better predictions. This may be viewed as data misuse, a violation of privacy, or a violation of the right to be forgotten.

Furthermore, there's an inherent ethical challenge in the goal of generating synthetic perspectives and opinions. The ability to synthetically generate opinions might inadvertently discourage practitioners from seeking real human input. This poses two primary risks: 1) it may lead to erroneous assumptions based on the synthetic data rather than actual human sentiments, and 2) it might marginalize authentic human participation, thereby weakening the quality and inclusivity of dataset and model development.

While our methods are designed to help detect when models may be incorrectly predicting human behaviors, they are most effective when applied to models performing imputation. Thus, advocating for the success of this work may inadvertently promote the creation and usage of models with the ethical concerns described above.

We urge creators of individualized prediction systems to always obtain consent from their users before applying models to their data and to maintain open and consistent communication about how their data may be used. We also advocate for a balanced approach, ensuring that while we progress in model development, real human perspectives remain at the core of our datasets and models.

References

[1] O. Sharir, B. Peleg, Y. Shoham, The Cost of Training NLP Models: A Concise Overview, 2020. URL: http://arxiv.org/abs/2004.08900, arXiv:2004.08900 [cs].

[2] A. Checco, K. Roitero, E. Maddalena, S. Mizzaro, G. Demartini, Let's agree to disagree: Fixing agreement measures for crowdsourcing, in: Fifth AAAI Conference on Human Computation and Crowdsourcing, 2017.

[3] S. Kairam, J. Heer, Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks, in: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 2016, pp. 1637–1648.

[4] A. Uma, D. Almanea, M. Poesio, Scaling and disagreements: Bias, noise, and ambiguity, Frontiers in Artificial Intelligence 5 (2022).

[5] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, M. Poesio, Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning, in: 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021.

[6] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, M. Poesio, SemEval-2023 Task 11: Learning With Disagreements (LeWiDi) (????).

[7] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110. URL: https://aclanthology.org/2022.tacl-1.6. doi:10.1162/tacl_a_00449, place: Cambridge, MA Publisher: MIT Press.

[8] S. Rendle, L. Schmidt-Thieme, Online-updating regularized kernel matrix factorization models for large-scale recommender systems, in: Proceedings of the 2008 ACM conference on Recommender systems, RecSys '08, Association for Computing Machinery, New York, NY, USA, 2008, pp. 251–258. URL: https://doi.org/10.1145/1454008.1454047. doi:10.1145/1454008.1454047.

[9] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural Collaborative Filtering, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, pp. 173–182. URL: https://doi.org/10.1145/3038912.3052569. doi:10.1145/3038912.3052569.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[11] M. Poesio, R. Artstein, The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account, in: Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky, 2005, pp. 76–83.

[12] Y. Versley, Vagueness and Referential Ambiguity in a Large-Scale Annotated Corpus, Research on

Language and Computation 6 (2008) 333–353. URL: https://doi.org/10.1007/s11168-008-9059-1. doi:10.1007/s11168-008-9059-1.

[13] R. Wan, K. Badillo-Urquiola, Dragonfly_captain at SemEval-2023 task 11: Unpacking disagreement with investigation of annotator demographics and task difficulty, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1978–1982. URL: https://aclanthology.org/2023.semeval-1.272. doi:10.18653/v1/2023.semeval-1.272.

[14] M. L. Gordon, M. S. Lam, J. S. Park, K. Patel, J. T. Hancock, T. Hashimoto, M. S. Bernstein, Jury Learning: Integrating Dissenting Voices into Machine Learning Models, 2022. URL: http://arxiv.org/abs/2202.02950. doi:10.1145/3491102.3502004, arXiv:2202.02950 [cs].

[15] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, M. S. Bernstein, The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 2021, pp. 1–14. URL: https://dl.acm.org/doi/10.1145/3411764.3445423. doi:10.1145/3411764.3445423.

[16] R. Wan, J. Kim, D. Kang, Everyone's voice matters: Quantifying annotation disagreement using demographic information, arXiv preprint arXiv:2301.05036 (2023).

[17] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from Disagreement: A Survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470. URL: https://www.jair.org/index.php/jair/article/view/12752. doi:10.1613/jair.1.12752.

[18] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems 22 (2004) 5–53. URL: https://dl.acm.org/doi/10.1145/963770.963772. doi:10.1145/963770.963772.

[19] F. O. Isinkaye, Y. O. Folajimi, B. A. Ojokoh, Recommendation systems: Principles, methods and evaluation, Egyptian Informatics Journal 16 (2015) 261–273. URL: https://www.sciencedirect.com/science/article/pii/S1110866515000341. doi:10.1016/j.eij.2015.06.005.

[20] Q.-V. Do, Matrix Factorization, 2022. URL: https://github.com/Quang-Vinh/matrix-factorization, original-date: 2020-06-04T00:10:11Z.

[21] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, Y. Choi, Social chemistry 101: Learning to reason about social and moral norms, ArXiv abs/2011.00620 (2020).

[22] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, ArXiv abs/1911.03891 (2020).

[23] B. Kennedy, M. Atari, A. Davani, L. Yeh, A. Omrani, Y. Kim, K. Coombs, S. Havaldar, G. Portillo-Wightman, E. Gonzalez, et al., Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale (2018).

[24] B. Kennedy, M. Atari, A. M. Davani, L. Yeh, A. Omrani, Y. Kim, K. Coombs, S. Havaldar, G. Portillo-Wightman, E. Gonzalez, et al., Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale, Language Resources and Evaluation 56 (2022) 79–108.

[25] M. Diaz, I. L. Johnson, A. Lazar, A. M. Piper, D. Gergle, Addressing age-related bias in sentiment analysis, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018).

[26] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, C. Potts, A computational approach to politeness with application to social factors, in: ACL, 2013.

[27] B. Roy, All About Missing Data Handling. Missing data is a every day problem… | by Baijayanta Roy | Towards Data Science, 2019. URL: https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184.

[28] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling Laws for Neural Language Models, 2020. URL: http://arxiv.org/abs/2001.08361. doi:10.48550/arXiv.2001.08361, arXiv:2001.08361 [cs, stat].

[29] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[30] H. F. , bert-base-uncased · Hugging Face, 2023. URL: https://huggingface.co/bert-base-uncased.

A. Multitask Model Details

The multitask model follows the specifications by [7]. Specifically, let $BERT$ be the Hugging Face "bert-base-uncased" model, which takes in a text, t_i, and outputs the embedding of the [CLS] token for that text [29, 30]. Then, let Lin represent a linear layer which takes in the embedding output by $BERT$ and outputs K values, where K is the number of valid annotation classes. We have M of these linear layers, one for each annotator j. Finally, let v_i be a single-dimensional array whose jth entry is 1 if the corresponding annotation $a_{i,j}$ is not missing (is valid), and 0 if it is missing (is not valid). Finally, let CE represent the cross entropy function of two vectors.

Then, the output of the model o_i for a given text t_i is computed as a single-dimensional array whose jth value is

$$o_{i,j} = Lin_j(BERT(t_i)).$$

And the loss for the model is computed as

$$CE(o_i \odot v_i, a_i).$$

Exact implementation details can be found in our code.

B. Dataset Details

Each dataset consists of two files: a text and annotation file. The text file consists of N texts, such that t_i refers to the ith text, where $1 \leq i \leq N$. The annotation file consists of annotations of text, and is a NxM matrix, where $a_{i,j}$ refers to the annotation given by the jth annotator for the ith text, where $1 \leq j \leq M$.

For all datasets, $a_{i,j}$ is an integer rating of the text. While different datasets have upper bounds of potential ratings, ratings which are numerically close to one another signify annotations which are semantically close to one another. In other words, for the datasets we use, a rating of 1 is similar to a rating of 2 and less similar to a rating of 5. This is in contrast to standard classification tasks, where class labels may differ significantly in semantics despite being close numerically.

C. Hyperparameters

C.1. NCF Matrix Factorization

The hyperparameters for NCF matrix factorization are

- Factors: [4, 8, 16, 32, 64, 128]
- Learning Rates: [0.001, 0.0005, 0.0001, 0.00005]

Hyperparameters are picked automatically through a grid search of all possible values during each run based on whichever hyperparameters achieve the lowest RMSE score on all of the training data.

C.2. Kernel Matrix Factorization

The hyperparameters for kernel matrix factorization are:

- Factors: [1, 2, 4, 8, 16, 32]
- Epochs: [1, 2, 4, 8, 16, 32, 64, 128, 256]
- Kernels: [linear, rbf, sigmoid]
- Gammas: (always set to auto)
- Regularization: [0.1, 0.01, 0.001]
- Learning Rate: [0.01, 0.001, 0.0001]
- Initial Mean: (always set to 0)
- Initial Standard Deviation: (always set to 0.1)
- Random Seed: [42 85]

The hyperparameters used for each imputation task are picked automatically based on a randomly-chosen held-out validation set consisting of 5% of the training data.

C.3. Multitask Model

The hyperparameters for the Multitask model are:

- Epochs: (always set to 10)
- Learning rate: (always set to 5e-5)

D. Additional Imputed Prompting Experiments

Method / Dataset	GHC	SBIC	SChem
Orig. Dist.	83.33%	76.67%	**50.00%**
NCF Dist.	83.33%	76.67%	**50.00%**
Maj. Voted	**88.89%**	**83.33%**	45.00%

Table 9
Accuracy of GPT-3 at making individualized predictions for a given text when provided with 1. The original distribution 2. The NCF-imputed distribution and 3. The majority-voted annotation for that text

Method / Dataset	GHC	SBIC	SChem
Not Imputed	**0.624**	**0.653**	**0.425**
Imputed	0.471	0.594	0.312

Table 10
Weighted F1 score of ChatGPT making individualized predictions for one out of three provided individuals. The highest F1 score for each dataset is in **bold**. Example prompts can be found in Appendix F.

In Table 9 we provide an overview of performance comparing the accuracy of GPT-3 for making individualized predictions when provided with either 1. The

Method / Dataset	GHC	SBIC	SChem
Not Imputed	**11.020**$\pm$**10.169**	10.525$\pm$10.484	**0.430**$\pm$**0.483**
Imputed	12.459$\pm$10.241	**8.921**$\pm$**9.687**	0.533$\pm$0.621

Table 11
KL divergence score (and standard deviation) of distributions predicted by ChatGPT compared to the true distributions for the given datasets. Imputation is done via the NCF method. Results are reported from the "no-context" prompt (see Appendix F).

original soft label 2. The imputed soft label or 3. The majority-voted label. Note that we expect a lower accuracy for SChem in comparison to SBIC or GHC since SChem has 5 labels, while SBIC and GHC have 3 and 2 labels respectively.

Interestingly, there was *no impact to accuracy* based on whether or not imputed versus original data was used. While Section 4.1 clearly indicates differences between imputed soft labels and original soft labels, GPT-3 appears to be robust to these differences when making individualized predictions.

We do see that providing the majority-voted annotation rather than the soft label improves performance by roughly 5% on GHC and 7% on SBIC. However, it also drops performance on SChem by 5%. This appears to indicate that for datasets with less labels, providing the majority-voted label enables GPT-3 to make better predictions than if one were to provide a soft label. However, as the number of labels increases, soft labels may provide more informative information for accurate predictions.

In Table 10 we display the impact of imputed data on making individualized predictions for one of three annotators whose data was provided in the prompt. The data clearly shows that imputation has a negative impact on ChatGPT's ability to make accurate individualized predictions.

Similar results are shown in Table 11 where we display the impact of imputed data on making soft label predictions. A high KL divergence score indicates a worse prediction; for GHC and SChem, imputation seems to harm the predictions, whereas for SBIC, imputation seems to help significantly. However, if we analyze the standard deviation, we see that it is often near if not greater than the mean, indicating a distribution that is skewed highly to the right, and suggesting that any changes in performance are not particularly significant.

E. Experiments to Support Low-Response Imputation

Overall, based on the data we have compiled into Table 12, there is no clear pattern for annotators with high response rate as to whether using imputed data rather than real data is more beneficial for making individualized predictions. We cannot test if this is the case on the annotators with a low-response-rate, as they do not have enough annotations to replace the imputed annotations.

Table 13 indicates that swapping the prompts may increase results in some cases, but, again, there~~'s~~ is no clear trend similar to the trend we saw for using imputed data, which can be verified again in this data by noticing that the imputed column consistently outperforms other columns for all datasets but Politeness.

Together, these two experiments show that the increase in F1 score is not due to the text before the prompt, and that it is the moderate increase in examples that imputation can provide, rather than the imputed data itself, that is likely the cause of the increased performance.

F. Imputed Prompting Prompt Details

F.1. Description of Prompts

For the highlighted ChatGPT experiment and ablation studies, each of the text portions was chosen from a list of possible options, and each possible combination of these options, along with multiple prompt versions, was used for an initial run on SBIC and politeness. After this initial run, the worst-performing prompts and prompt options were removed, and all datasets were run again. The results reported are the best results among all prompts used. Exact details, including all of the full prompts, prompt options, examples, and outputs can be found in our code.

For GPT-3, we provide either the true (original/non-imputed) soft label, the imputed soft label, or the true majority-voted (aggregate) label for the target text. For the distributional label, we ignore the annotator's actual label when computing the distribution, so as not to cause data leakage. However, when computing the original majority-voted annotation, we leave in the annotator's label for the target example. For the non-highlighted ChatGPT experiments, when making soft label predictions we use the soft label from the imputed data, rather than real data. When making individualized annotation predictions, the example shots are chosen to differ from the original such that the imputed annotation can be used.

Version	Replaced with Original			Standard			
Prompt	"Imputed"	Original	Combined	Imputed	Original	Combined	Low30
Politeness	**0.213**	0.186	**0.199**	0.094	0.186	0.085	0.14
GHC	**0.896**	0.846	0.745	0.858	0.846	**0.796**	0.86
SChem	0.604	**0.359**	0.563	**0.615**	0.355	**0.620**	0.56
SChem5L	0.578	0.581	**0.492**	**0.595**	0.581	0.407	0.65
SBIC	0.820	0.733	**0.820**	**0.831**	0.733	0.760	0.95
Sentiment	**0.513**	0.063	0.513	0.496	**0.139**	**0.558**	0.60

Table 12
Table of F1 scores measuring individualized predictions made by ChatGPT, given data from *high-response-rate* annotators. In the "Replaced with Original" condition, imputed data is replaced with original data (but the rest of the prompt remains the same). In the "Standard" condition, imputed data remains imputed. The "Low30" section copies over the highest F1 score from from Table 8, which uses data from low-response-rate annotators, for direct comparison to these results. Scores are listed in bold if they outcompete their "Replaced with Original" or "Standard" counterpart. F1 scores from the "Low30" column are underlined if they outperform all high-response-rate scores.

Version	Original Prompt			Swapped Prompt		
Prompt	Imputed	Original	Combined	Imputed	Original	Combined
Politeness	0.033	0.050	0.027	**0.067**	**0.144**	**0.128**
GHC	0.858	0.745	**0.846**	0.858	**0.846**	0.796
SChem	0.592	0.502	**0.532**	0.592	0.502	0.502
SChem5L	**0.747**	0.493	**0.697**	0.646	**0.519**	0.600
SBIC	0.952	0.926	0.932	0.952	0.926	0.932
Sentiment	0.600	0.059	0.585	**0.604**	**0.31**	**0.638**

Table 13
F1 Scores of ChatGPT predicting annotations of low-response-rate individuals, but with the text preceding the imputed and original examples swapped. F1 scores are **bolded** if they are higher than their swapped or original counterpart , and underlined if the data (imputed, original, or combined) used in the prompt outperforms other data within the same condition (original or swapped).

F.2. Prompt Skeletons

This section displays the skeletons of each of the prompts used. In practice, the portions of the skeleton surrounded by curly braces are replaced with data, which can be seen in Section F.3.

F.2.1. Highlighted ChatGPT Original Data Skeleton Prompt 1

```
{dataset_description}

{orig_examples_header}
{orig_examples}

{target_example_header}
{target_example}
{final_words}
```

F.2.2. Highlighted ChatGPT Original Data Skeleton Prompt 2

```
{dataset_description}

{target_example_header}
{target_example}
{final_words}
```

F.2.3. Highlighted ChatGPT Original Data Skeleton Prompt 3

```
{dataset_description}

{instructions}
{target_example_header}
{target_example}
{final_words}
```

F.2.4. Highlighted ChatGPT Combined Data Skeleton Prompt

```
{imputed_examples_header}
{imputed_examples}

{orig_examples_header}
{orig_examples}

{target_example_header}
{target_example}
```

F.2.5. Highlighted ChatGPT Imputed Data Skeleton Prompt 1

```
{imputed_examples}

{target_example}
```

F.2.6. Highlighted ChatGPT Imputed Data Skeleton Prompt 2

```
{imputed_examples_header}
{imputed_examples}

{target_example_header}
{target_example}
```

F.2.7. GPT-3 Distributional Skeleton Prompt

```
Here's a description of a dataset:
{dataset_description}

Given the previous dataset description, your goal is to predict how one of the annotators of the previous dataset would annotate an example from that dataset. You will be given {n_shots} samples of how that particular annotator has responded to other examples and be shown the distributional label of how all annotators have annotated the target example, and will then complete the prediction for the target example as that annotator would.

Here's the samples of how the particular annotator has responded to other examples:
{shots}

Here's how the distributional label of how all annotators have annotated the target example:
{other_shots}

How would the particular annotator annotate the target example?
{target_example_line}
ANSWER:
```

F.2.8. GPT-3 Individual Skeleton Prompt

```
Here's a description of a dataset:
{dataset_description}

Given the previous dataset description, your goal is to predict how one of the annotators of the previous dataset would annotate an example from that dataset. You will be given {n_shots} samples of how that particular annotator has responded to other examples and {k_shots} sample of how others have annotated the target example, and will then complete the prediction for the target example as that annotator would.

Here's the samples of how the particular annotator has responded to other examples:
{shots}

Here's the samples of how others have annotated the target example:
{other_shots}

How would the particular annotator annotate the target example?
{target_example_line}
ANSWER:
```

F.2.9. GPT-3 Majority-Voted Skeleton Prompt

```
Here's a description of a dataset:
{dataset_description}

Given the previous dataset description, your goal is to predict how one of the annotators of the previous dataset would annotate an example from that dataset. You will be given {n_shots} samples of how that particular annotator has responded to other examples and be shown what the plurality of annotators gave as a label, and will then complete the prediction for the target example as that annotator would.

Here's the samples of how the particular annotator has responded to other examples:
{shots}

Here's how the plurality of annotators labeled the target example:
{other_shots}

How would the particular annotator annotate the target example?
{target_example_line}
ANSWER:
```

F.2.10. ChatGPT Soft Label Skeleton Prompt

```
{soft_label_examples}
{prediction_text}
```

F.2.11. (Unused) ChatGPT Contextual Soft Label Skeleton

```
Here is a description of a dataset:
{dataset_description}

Your goal is to predict the soft label given
by the raters on a particular text.

Here are a few examples of texts and their
soft label:
{soft_label_examples}

Now, you will make your prediction (if you
are unsure, just give your best estimate):
{prediction_text}
```

F.2.12. ChatGPT One of Three Individualized Skeleton Prompt

```
Description of the dataset:
{dataset_description}

How annotator A has labeled some examples:
{annotator_A_examples}

How annotator B has labeled some examples:
{annotator_B_examples}

How annotator C has labeled some examples:
{annotator_C_examples}

How might annotator {target_annotator}
label this example? Output the annotation
that seems most likely.
{target_example}
```

F.3. Full Prompts

This section demonstrates how the skeletons are completed with full data for the prompts that are actually sent to the OpenAI GPT models. Version numbers such as "v4.-1.0.-1.1" indicate which fillers were used for the prompt. The value "v4.-1.0.-1.1" indicates that the 5th option was used for the first filler, the second filler was not included in the prompt, the first option was used for the third filler, the second to last filler was not included, and the 2nd option was used for the last filler.

The fillers, in order, are:

```
1. orig_examples_header
2. imputed_examples_header
3. target_example_header
4. instructions
5. final_words
```

As an example, the options for the filler may look like:

```
Fillers for final_words:
1. "Your output should be a single integer
corresponding to the label."
2. "Your output should be a single integer
and nothing else."
3. "The only valid output is a single
integer."
4. "If you output anything other than
a single integer, your output will be
considered invalid."
5. "If you output anything other than a
single integer, your output will harm the
integrity of our dataset."
6. "If you output anything other than
a single integer (and absolutely nothing
else, including explanatory text), your
output will invalidate the dataset."
7. "If you output anything other than
a single integer (and absolutely nothing
else, including explanatory text), your
output will invalidate the dataset. So,
please only output a single integer."
```

Explicit details of all filler options for all experiments can be found in the code.

F.3.1. Highlighted ChatGPT Original Data Full Prompt 1 (v4.-1.0.-1.1)

```
In this dataset, people are given rules
of thumb, such as "You should be accepting
of other people's bodies." People then rate
the rule of thumb as to how much of the world
population they think would agree with that
rule of thumb. 4 means "universal". 3 means
"common". 2 means "controversial". 1 means
"occasional". 0 means "almost no one".

Examples from the dataset (there may not be
any):Examples from the dataset (there may
not be any)
Example 1:
Text: It's good to trust those you care
about.
Annotation from annotator: 3

Example 2:
Text: It's wrong to grow apart from your
partner.
Annotation from annotator: 1

Example 3:
Text: It's good to find proof of cheating
you perceive to be happening.
Annotation from annotator: 3
```

Example 4:
Text: It’s good to get a friend’s help.
Annotation from annotator: 3

Example 5:
Text: You should always help out your customers when they have a problem.
Annotation from annotator: 3

Example 6:
Text: You shouldn’t post inappropriate pictures online.
Annotation from annotator: 3

Example 7:
Text: You shoulld not worry about others being jealous when posting things on facebook.
Annotation from annotator: 3

Example 8:
Text: You should explain to a friend why you cannot give them money if you cannot.
Annotation from annotator: 2

Example 9:
Text: You shouldn’t video tape someone without their permission.
Annotation from annotator: 3

Example 10:
Text: It’s good to give advice to your friends.
Annotation from annotator: 3

Example 11:
Text: You shouldn’t judge people by how they look
Annotation from annotator: 3

Example 12:
Text: It is okay to show emotions.
Annotation from annotator: 2

Example 13:
Text: It’s bad to hate babies.
Annotation from annotator: 3

Example 14:
Text: You shouldn’t scare your coworkers.
Annotation from annotator: 3

Example 15:
Text: It’s not okay to rely on technology for health.
Annotation from annotator: 1

Example 16:
Text: You can’t expect love young to last forever.
Annotation from annotator: 3

Example 17:
Text: It’s good to want to do the things it takes to start a grown up life.
Annotation from annotator: 3

Example 18:
Text: It’s normal for married couples to have sexual relations.
Annotation from annotator: 4

Example 19:
Text: You should always know why you do things.
Annotation from annotator: 3

Example 20:
Text: It’s understandable to want to get back at someone.
Annotation from annotator: 3

Example 21:
Text: You shouldn’t date a bad partner.
Annotation from annotator: 3

Example 22:
Text: It’s ok to make your own reproductive choices.
Annotation from annotator: 2

Example 23:
Text: It’s good for a parent to want their children to make the best decisions possible.
Annotation from annotator: 3

Example 24:
Text: It is distracting to pin everyone in an admin team on a discord server.
Annotation from annotator: 3

Example 25:
Text: It is okay to not share something that you did a lonf time ago.
Annotation from annotator: 3

Example 26:
Text: It is good to be yourself.

Annotation from annotator: 3

Example 27:
Text: It's expected that you'll stay home from work when you are sick.
Annotation from annotator: 3

Example 28:
Text: It's inappropriate for a parent to help a teenage boy shower.
Annotation from annotator: 3

Example 29:
Text: It's good to want to be able to back up your words.
Annotation from annotator: 4

Example 30:
Text: You should treat employees well
Annotation from annotator: 3

Target Example:
Text: It's OK not to take life too seriously.
Annotation from annotator:
Your output should be a single integer and nothing else.

F.3.2. Highlighted ChatGPT Original Data Full Prompt 2 (v-1.-1.0.-1.0)

In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

Target Example:
Text: People shouldn't betray the trust of those they're close to.
Annotation from annotator:
Your output should be a single integer corresponding to the label.

F.3.3. Highlighted ChatGPT Original Data Full Prompt 3 (v-1.-1.0.1.0)

In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

Now you will make your prediction (if you are unsure, just give your best estimate.)
Your output should be an integer label:
Target Example:
Text: People shouldn't betray the trust of those they're close to.
Annotation from annotator:
Your output should be a single integer corresponding to the label.

F.3.4. Highlighted ChatGPT Combined Full Prompt (v1.0.0.-1.-1

Estimated Examples:
Example 1:
Text: It's wise to never let yourself get embarrassed by your significant other.
Annotation from annotator: 2

Example 2:
Text: It's not normal to dislike everybody
Annotation from annotator: 3

Example 3:
Text: It's rude to dislike your best friend's boyfriend.
Annotation from annotator: 2

Example 4:
Text: It's okay to be upset if someone you report to at work lets you take the blame for something you didn't do.
Annotation from annotator: 3

Example 5:
Text: You should always respect your partner's choices.
Annotation from annotator: 3

Example 6:
Text: It's wrong to spend other people's money.
Annotation from annotator: 3

Example 7:
Text: You shouldn't be so scared about getting married.
Annotation from annotator: 2

Example 8:
Text: It's wrong to want to hide who you

really are.
Annotation from annotator: 2

Example 9:
Text: You shouldn't have sex with people if you have herpes.
Annotation from annotator: 3

Example 10:
Text: You shouldn't video tape someone without their permission.
Annotation from annotator: 3

Example 11:
Text: You should thank someone when they invite tou to thier wedding.
Annotation from annotator: 3

Example 12:
Text: You can't expect your family members to support you if you're lazy.
Annotation from annotator: 3

Example 13:
Text: It's expected that revealing secrets will lead to trouble.
Annotation from annotator: 3

Example 14:
Text: If you are uncomfortable with your partner feeding your kids, you should not let your partner watch your children.
Annotation from annotator: 3

Example 15:
Text: It's good to want to keep your children safe in life.
Annotation from annotator: 3

Example 16:
Text: It's okay to have sex on camera.
Annotation from annotator: 2

Example 17:
Text: It's wrong for partners to lie to each other.
Annotation from annotator: 3

Example 18:
Text: It is reasonable to want to hold onto loving memories after losing someone.
Annotation from annotator: 3

Example 19:
Text: You shouldn't play pranks on people.
Annotation from annotator: 2

Example 20:
Text: It's okay to have an abortion.
Annotation from annotator: 2

Example 21:
Text: It's wrong to think you're responsible for somebody else's life
Annotation from annotator: 2

Example 22:
Text: You should not get angry when your chronically ill relative needs help.
Annotation from annotator: 3

Example 23:
Text: You shouldn't be bitter and salty about petty things.
Annotation from annotator: 3

Example 24:
Text: It's good to play games of chance where you can win money.
Annotation from annotator: 2

Example 25:
Text: Relationships are expected to have clear communication.
Annotation from annotator: 3

Example 26:
Text: You shouldn't try to get sympathy from people just because you're in the hospital.
Annotation from annotator: 2

Example 27:
Text: You shouldn't be friends with someone that talks badly about you behind your back.
Annotation from annotator: 3

Example 28:
Text: It's okay to carry your pet in public as a status symbol
Annotation from annotator: 1

Example 29:
Text: It's not okay to be in a romantic relationship with a teacher.
Annotation from annotator: 3

Example 30:
Text: You can't expect everyone to like you.
Annotation from annotator: 3

Human-rated Examples:
Example 1:
Text: It's good to be honest when you give your kids the sex talk.
Annotation from annotator: 3

Example 2:
Text: It's wrong for people that care about each other to argue.
Annotation from annotator: 2

Example 3:
Text: It's inappropriate to make passes at an underling.
Annotation from annotator: 3

Example 4:
Text: Couple should work together to solve problems
Annotation from annotator: 3

Example 5:
Text: It's okay to throw away compostables.
Annotation from annotator: 2

Example 6:
Text: It's bad to smoke.
Annotation from annotator: 3

Example 7:
Text: You can't expect someone to pay off their loans immediately.
Annotation from annotator: 3

Example 8:
Text: It's not necessary to be strong
Annotation from annotator: 3

Example 9:
Text: You shouldn't put up an antenna if it's snowing outside.
Annotation from annotator: 1

Example 10:
Text: It's good to have a best friend.
Annotation from annotator: 3

Example 11:
Text: It's bad to lose track of your belongings.
Annotation from annotator: 3

Example 12:
Text: It's not okay to have to rely on special tools when they aren't really needed.
Annotation from annotator: 2

Target Example:
Text: People shouldn't betray the trust of those they're close to.
Annotation from annotator:

F.3.5. Highlighted ChatGPT Imputed Data Full Prompt 1 (v-1.-1.-1.-1.-1)

Example 1:
Text: It's wise to never let yourself get embarrassed by your significant other.
Annotation from annotator: 2

Example 2:
Text: It's not normal to dislike everybody
Annotation from annotator: 3

Example 3:
Text: It's rude to dislike your best friend's boyfriend.
Annotation from annotator: 2

Example 4:
Text: It's okay to be upset if someone you report to at work lets you take the blame for something you didn't do.
Annotation from annotator: 3

Example 5:
Text: You should always respect your partner's choices.
Annotation from annotator: 3

Example 6:
Text: It's wrong to spend other people's money.
Annotation from annotator: 3

Example 7:
Text: You shouldn't be so scared about getting married.
Annotation from annotator: 2

Example 8:
Text: It's wrong to want to hide who you really are.
Annotation from annotator: 2

Example 9:
Text: You shouldn't have sex with people if you have herpes.

Annotation from annotator: 3

Example 10:
Text: You shouldn't video tape someone without their permission.
Annotation from annotator: 3

Example 11:
Text: You should thank someone when they invite tou to thier wedding.
Annotation from annotator: 3

Example 12:
Text: You can't expect your family members to support you if you're lazy.
Annotation from annotator: 3

Example 13:
Text: It's expected that revealing secrets will lead to trouble.
Annotation from annotator: 3

Example 14:
Text: If you are uncomfortable with your partner feeding your kids, you should not let your partner watch your children.
Annotation from annotator: 3

Example 15:
Text: It's good to want to keep your children safe in life.
Annotation from annotator: 3

Example 16:
Text: It's okay to have sex on camera.
Annotation from annotator: 2

Example 17:
Text: It's wrong for partners to lie to each other.
Annotation from annotator: 3

Example 18:
Text: It is reasonable to want to hold onto loving memories after losing someone.
Annotation from annotator: 3

Example 19:
Text: You shouldn't play pranks on people.
Annotation from annotator: 2

Example 20:
Text: It's okay to have an abortion.
Annotation from annotator: 2

Example 21:
Text: It's wrong to think you're responsible for somebody else's life
Annotation from annotator: 2

Example 22:
Text: You should not get angry when your chronically ill relative needs help.
Annotation from annotator: 3

Example 23:
Text: You shouldn't be bitter and salty about petty things.
Annotation from annotator: 3

Example 24:
Text: It's good to play games of chance where you can win money.
Annotation from annotator: 2

Example 25:
Text: Relationships are expected to have clear communication.
Annotation from annotator: 3

Example 26:
Text: You shouldn't try to get sympathy from people just because you're in the hospital.
Annotation from annotator: 2

Example 27:
Text: You shouldn't be friends with someone that talks badly about you behind your back.
Annotation from annotator: 3

Example 28:
Text: It's okay to carry your pet in public as a status symbol
Annotation from annotator: 1

Example 29:
Text: It's not okay to be in a romantic relationship with a teacher.
Annotation from annotator: 3

Example 30:
Text: You can't expect everyone to like you.
Annotation from annotator: 3

Text: People shouldn't betray the trust of those they're close to.
Annotation from annotator:

F.3.6. Highlighted ChatGPT Imputed Data Full Prompt 2 (v-1.0.0.-1.-1)

```
Estimated Examples:
Example 1:
Text: It's wise to never let yourself get embarrassed by your significant other.
Annotation from annotator: 2

Example 2:
Text: It's not normal to dislike everybody
Annotation from annotator: 3

Example 3:
Text: It's rude to dislike your best friend's boyfriend.
Annotation from annotator: 2

Example 4:
Text: It's okay to be upset if someone you report to at work lets you take the blame for something you didn't do.
Annotation from annotator: 3

Example 5:
Text: You should always respect your partner's choices.
Annotation from annotator: 3

Example 6:
Text: It's wrong to spend other people's money.
Annotation from annotator: 3

Example 7:
Text: You shouldn't be so scared about getting married.
Annotation from annotator: 2

Example 8:
Text: It's wrong to want to hide who you really are.
Annotation from annotator: 2

Example 9:
Text: You shouldn't have sex with people if you have herpes.
Annotation from annotator: 3

Example 10:
Text: You shouldn't video tape someone without their permission.
Annotation from annotator: 3

Example 11:
Text: You should thank someone when they invite tou to thier wedding.
Annotation from annotator: 3

Example 12:
Text: You can't expect your family members to support you if you're lazy.
Annotation from annotator: 3

Example 13:
Text: It's expected that revealing secrets will lead to trouble.
Annotation from annotator: 3

Example 14:
Text: If you are uncomfortable with your partner feeding your kids, you should not let your partner watch your children.
Annotation from annotator: 3

Example 15:
Text: It's good to want to keep your children safe in life.
Annotation from annotator: 3

Example 16:
Text: It's okay to have sex on camera.
Annotation from annotator: 2

Example 17:
Text: It's wrong for partners to lie to each other.
Annotation from annotator: 3

Example 18:
Text: It is reasonable to want to hold onto loving memories after losing someone.
Annotation from annotator: 3

Example 19:
Text: You shouldn't play pranks on people.
Annotation from annotator: 2

Example 20:
Text: It's okay to have an abortion.
Annotation from annotator: 2

Example 21:
Text: It's wrong to think you're responsible for somebody else's life
Annotation from annotator: 2

Example 22:
Text: You should not get angry when your chronically ill relative needs help.
```

Annotation from annotator: 3

Example 23:
Text: You shouldn't be bitter and salty about petty things.
Annotation from annotator: 3

Example 24:
Text: It's good to play games of chance where you can win money.
Annotation from annotator: 2

Example 25:
Text: Relationships are expected to have clear communication.
Annotation from annotator: 3

Example 26:
Text: You shouldn't try to get sympathy from people just because you're in the hospital.
Annotation from annotator: 2

Example 27:
Text: You shouldn't be friends with someone that talks badly about you behind your back.
Annotation from annotator: 3

Example 28:
Text: It's okay to carry your pet in public as a status symbol
Annotation from annotator: 1

Example 29:
Text: It's not okay to be in a romantic relationship with a teacher.
Annotation from annotator: 3

Example 30:
Text: You can't expect everyone to like you.
Annotation from annotator: 3

Target Example:
Text: People shouldn't betray the trust of those they're close to.
Annotation from annotator:

F.3.7. GPT-3 Non-imputed Distributional Full Prompt

Here's a description of a dataset:
In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

Given the previous dataset description, your goal is to predict how one of the annotators of the previous dataset would annotate an example from that dataset. You will be given 4 samples of how that particular annotator has responded to other examples and be shown the distributional label of how all annotators have annotated the target example, and will then complete the prediction for the target example as that annotator would.

Here's the samples of how the particular annotator has responded to other examples:
1.
EXAMPLE: It's good to be honest when you give your kids the sex talk.
ANSWER: 4

2.
EXAMPLE: It's good to want others to celebrate your joys with you.
ANSWER: 3

3.
EXAMPLE: It is dangerous for a doctor to misdiagnose a patient.
ANSWER: 4

4.
EXAMPLE: It's bad for a child to be in danger.
ANSWER: 4

Here's how the distributional label of how all annotators have annotated the target example:
0.00% of people responded with 0
18.00% of people responded with 1
76.00% of people responded with 2
6.00% of people responded with 3

How would the particular annotator annotate the target example?
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER:

F.3.8. GPT-3 Imputed Distributional Full Prompt

Here's a description of a dataset:
In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

Given the previous dataset description, your goal is to predict how one of the annotators of the previous dataset would annotate an example from that dataset. You will be given 4 samples of how that particular annotator has responded to other examples and be shown the distributional label of how all annotators have annotated the target example, and will then complete the prediction for the target example as that annotator would.

Here's the samples of how the particular annotator has responded to other examples:
1.
EXAMPLE: It's good to be honest when you give your kids the sex talk.
ANSWER: 4

2.
EXAMPLE: It's good to want others to celebrate your joys with you.
ANSWER: 3

3.
EXAMPLE: It is dangerous for a doctor to misdiagnose a patient.
ANSWER: 4

4.
EXAMPLE: It's bad for a child to be in danger.
ANSWER: 4

Here's how the distributional label of how all annotators have annotated the target example:
0.00% of people responded with 0
10.00% of people responded with 1
85.00% of people responded with 2
5.00% of people responded with 3

How would the particular annotator annotate the target example?
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER:

F.3.9. GPT-3 Individual Full Prompt

Here's a description of a dataset:
In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

Given the previous dataset description, your goal is to predict how one of the annotators of the previous dataset would annotate an example from that dataset. You will be given 4 samples of how that particular annotator has responded to other examples and 49 sample of how others have annotated the target example, and will then complete the prediction for the target example as that annotator would.

Here's the samples of how the particular annotator has responded to other examples:
1.
EXAMPLE: It's good to be honest when you give your kids the sex talk.
ANSWER: 4

2.
EXAMPLE: It's good to want others to celebrate your joys with you.
ANSWER: 3

3.
EXAMPLE: It is dangerous for a doctor to misdiagnose a patient.
ANSWER: 4

4.
EXAMPLE: It's bad for a child to be in danger.
ANSWER: 4

Here's the samples of how others have annotated the target example:
1.
EXAMPLE: It's expected that families will not get along well with their in-laws.

ANSWER: 1

2.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

3.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

4.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

5.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

6.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

7.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 3

8.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 1

9.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 3

10.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 1

11.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 1

12.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

13.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

14.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

15.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

16.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 1

17.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

18.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

19.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

20.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

21.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

22.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

23.
EXAMPLE: It’s expected that families will not get along well with their in-laws.

ANSWER: 2

24.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

25.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

26.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

27.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

28.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

29.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

30.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

31.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

32.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 3

33.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

34.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

35.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

36.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

37.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

38.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

39.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

40.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

41.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 1

42.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

43.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 2

44.
EXAMPLE: It’s expected that families will not get along well with their in-laws.
ANSWER: 1

45.
EXAMPLE: It’s expected that families will not get along well with their in-laws.

ANSWER: 2

46.
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER: 2

47.
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER: 2

48.
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER: 1

49.
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER: 2

How would the particular annotator annotate the target example?
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER:

F.3.10. GPT-3 Majority-Voted Full Prompt

Here's a description of a dataset:
In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

Given the previous dataset description, your goal is to predict how one of the annotators of the previous dataset would annotate an example from that dataset. You will be given 4 samples of how that particular annotator has responded to other examples and be shown what the plurality of annotators gave as a label, and will then complete the prediction for the target example as that annotator would.

Here's the samples of how the particular annotator has responded to other examples:
1.
EXAMPLE: It's good to be honest when you give your kids the sex talk.
ANSWER: 4

2.
EXAMPLE: It's good to want others to celebrate your joys with you.
ANSWER: 3

3.
EXAMPLE: It is dangerous for a doctor to misdiagnose a patient.
ANSWER: 4

4.
EXAMPLE: It's bad for a child to be in danger.
ANSWER: 4

Here's how the plurality of annotators labeled the target example:
1.
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER: 2

How would the particular annotator annotate the target example?
EXAMPLE: It's expected that families will not get along well with their in-laws.
ANSWER:

F.3.11. ChatGPT Non-imputed Soft Label Full Prompt

Example 390
Text: It's hurtful for a parent to enroll a child in classes without permission.
Soft labels:
0.00% of responsive annotators labeled the text with 0
14.00% of responsive annotators labeled the text with 1
54.00% of responsive annotators labeled the text with 2
28.00% of responsive annotators labeled the text with 3
4.00% of responsive annotators labeled the text with 4

Example 261
Text: It's okay to enjoy powdered milk.
Soft labels:
0.00% of responsive annotators labeled the text with 0
2.00% of responsive annotators labeled the

text with 1
16.00% of responsive annotators labeled the text with 2
52.00% of responsive annotators labeled the text with 3
30.00% of responsive annotators labeled the text with 4

Example 20
Text: It's good to find proof of cheating you perceive to be happening.
Soft labels:
0.00% of responsive annotators labeled the text with 0
0.00% of responsive annotators labeled the text with 1
16.00% of responsive annotators labeled the text with 2
64.00% of responsive annotators labeled the text with 3
20.00% of responsive annotators labeled the text with 4

Example 84
Text: People shouldn't betray the trust of those they're close to.
Soft labels:
0.00% of responsive annotators labeled the text with 0
0.00% of responsive annotators labeled the text with 1
0.00% of responsive annotators labeled the text with 2
44.00% of responsive annotators labeled the text with 3
56.00% of responsive annotators labeled the text with 4
Target Text: You should always question someone's character when they say something inappropriate.
Soft labels:

F.3.12. ChatGPT Imputed Soft Label Full Prompt

Example 390
Text: It's hurtful for a parent to enroll a child in classes without permission.
Soft labels:
0.00% of responsive annotators labeled the text with 0
7.00% of responsive annotators labeled the text with 1
67.00% of responsive annotators labeled the text with 2
24.00% of responsive annotators labeled the text with 3
2.00% of responsive annotators labeled the text with 4

Example 261
Text: It's okay to enjoy powdered milk.
Soft labels:
0.00% of responsive annotators labeled the text with 0
1.00% of responsive annotators labeled the text with 1
8.00% of responsive annotators labeled the text with 2
64.00% of responsive annotators labeled the text with 3
27.00% of responsive annotators labeled the text with 4

Example 20
Text: It's good to find proof of cheating you perceive to be happening.
Soft labels:
0.00% of responsive annotators labeled the text with 0
0.00% of responsive annotators labeled the text with 1
8.00% of responsive annotators labeled the text with 2
78.00% of responsive annotators labeled the text with 3
14.00% of responsive annotators labeled the text with 4

Example 84
Text: People shouldn't betray the trust of those they're close to.
Soft labels:
0.00% of responsive annotators labeled the text with 0
0.00% of responsive annotators labeled the text with 1
0.00% of responsive annotators labeled the text with 2
47.00% of responsive annotators labeled the text with 3
53.00% of responsive annotators labeled the text with 4
Target Text: You should always question someone's character when they say something inappropriate.
Soft labels:

F.3.13. ChatGPT Non-imputed One of Three Full Prompt

```
Description of the dataset:
In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

How annotator A has labeled some examples:
Example 1:
Text: You should be focused on learning while at school.
Annotation from annotator: 3.0

Example 2:
Text: It is dangerous for a doctor to misdiagnose a patient.
Annotation from annotator: 3.0

Example 3:
Text: It's understandable if you can't keep your political leanings out of your religion lectures.
Annotation from annotator: 3.0

Example 4:
Text: You shouldn't date a bad partner.
Annotation from annotator: 3.0

How annotator B has labeled some examples:
Example 1:
Text: You shouldn't be with someone who sleeps with your kids.
Annotation from annotator: 4.0

Example 2:
Text: You shouldn't have sex with people if you have herpes.
Annotation from annotator: 3.0

Example 3:
Text: You shouldn't be crazy about things.
Annotation from annotator: 2.0

Example 4:
Text: If you don't want to crash, you shouldn't drive recklessly.
Annotation from annotator: 3.0

How annotator C has labeled some examples:
Example 1:
Text: You should only call the cops for emergencies.
Annotation from annotator: 3.0

Example 2:
Text: It is dangerous for a grandparent to risk a fall.
Annotation from annotator: 3.0

Example 3:
Text: Employers are expected to let employees go when business is down.
Annotation from annotator: 4.0

Example 4:
Text: People shouldn't assume it's okay to take something just because it's not nailed down.
Annotation from annotator: 4.0

How might annotator A label this example? Output the annotation that seems most likely.
Text: You should always question someone's character when they say something inappropriate.
Annotation from annotator:
```

F.3.14. ChatGPT Imputed One of Three Full Prompt

```
Description of the dataset:
In this dataset, people are given rules of thumb, such as "You should be accepting of other people's bodies." People then rate the rule of thumb as to how much of the world population they think would agree with that rule of thumb. 4 means "universal". 3 means "common". 2 means "controversial". 1 means "occasional". 0 means "almost no one".

How annotator A has labeled some examples:
Example 1:
Text: It's hurtful for  a parent to enroll a child in classes without permission.
Annotation from annotator: 2.0

Example 2:
Text: It's okay to enjoy powdered milk.
Annotation from annotator: 3.0

Example 3:
Text: It's good to find proof of cheating you perceive to be happening.
```

```
Annotation from annotator: 3.0

Example 4:
Text: People shouldn't betray the trust of
those they're close to.
Annotation from annotator: 3.0

How annotator B has labeled some examples:
Example 1:
Text: It's hurtful for  a parent to enroll
a child in classes without permission.
Annotation from annotator: 2.0

Example 2:
Text: It's okay to enjoy powdered milk.
Annotation from annotator: 3.0

Example 3:
Text: It's good to find proof of cheating
you perceive to be happening.
Annotation from annotator: 3.0

Example 4:
Text: People shouldn't betray the trust of
those they're close to.
Annotation from annotator: 3.0

How annotator C has labeled some examples:
Example 1:
Text: It's hurtful for  a parent to enroll
a child in classes without permission.
Annotation from annotator: 2.0

Example 2:
Text: It's okay to enjoy powdered milk.
Annotation from annotator: 2.0

Example 3:
Text: It's good to find proof of cheating
you perceive to be happening.
Annotation from annotator: 3.0

Example 4:
Text: People shouldn't betray the trust of
those they're close to.
Annotation from annotator: 4.0

How might annotator A label this example?
Output the annotation that seems most
likely.
Text: You should always question someone's
character when they say something
inappropriate.
Annotation from annotator:
```

G. PCA Results

One of the fundamental aspects of imputation methods is how they treat and interpret data. In the 2-dimensional scatter plot based on the first two principal components of imputed datasets, clear variations can be observed across different imputation techniques. This visualization underscores the unique characteristics of each imputation method. Refer to Figure 6 for a detailed comparison.

H. Variance and Disagreement

Post-imputation, a notable observation is the drop in variance, as shown in Figure 7. This phenomenon can be attributed to the fact that most imputation methods tend to approximate missing values based on observed patterns in the data, leading to a convergence of values around certain estimates.

I. Soft Label Analysis Extra Examples

Our code to generate the full websites containing all of the examples is publicly available. Here, in Figures 8, 9, 10, 11, 12, and 13, we provide a subset of examples demonstrating high and low KL divergence scores from each of the datasets.

J. Disagreement Levels

The computation to determine whether an example has is "low", "medium", or "high" disagreement was done individually for each fold of the data. When given a fold of data, we first compute the proportion of people who disagreed with the majority-voted label. (Note that ties in the majority-voted label do not impact this computation, since the same number of people will disagree regardless of which label is chosen among the tied options.) Then, we assign a threshold for "low" and "high" disagreement: any examples with disagreement equal to or lower than the "low" threshold are considered to have "low" disagreement, while any examples with disagreement equal to or greater than the "high" threshold are considered to have "high" disagreement. The number of examples in each category is a sum across all five folds of that dataset of examples that matched the threshold for that category.

The choice of thresholds must satisfy three rules: (1) The high threshold must be higher than the low threshold (2) There must be at least some examples in each category (3) The variance among the number of examples in each category must be minimized. When looking at Table 7, it may seem odd that the number of examples in each category is so varied, given the explicit minimization of

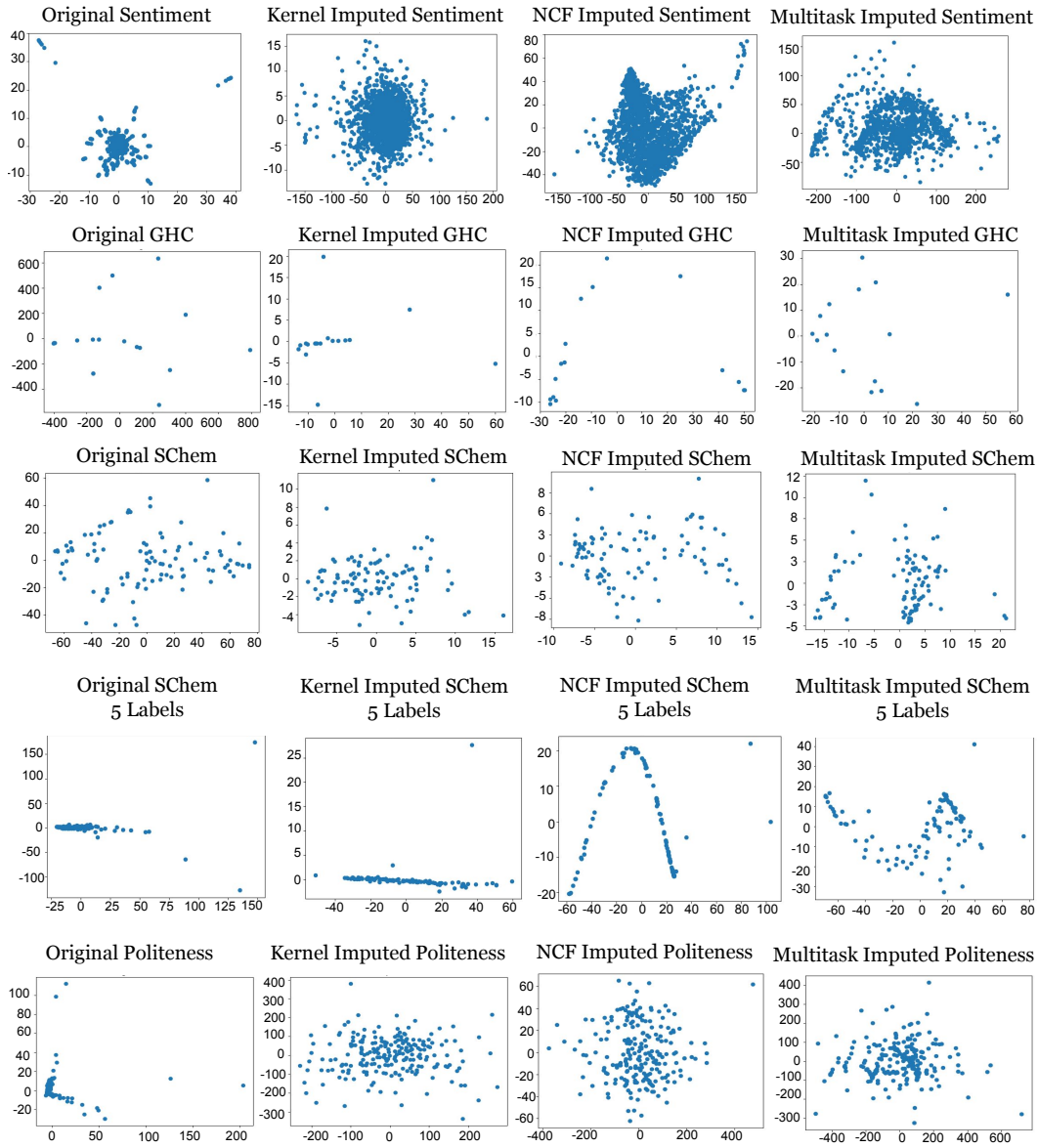

Figure 6: PCA projections of each of the datasets after different forms of imputation.

variance in the rules. However, this occurs because there are many examples that have the exact same level of disagreement; rather than split these examples into two different categories, we opted to ensure that examples with the same level of disagreement were always labeled with the same level of disagreement.

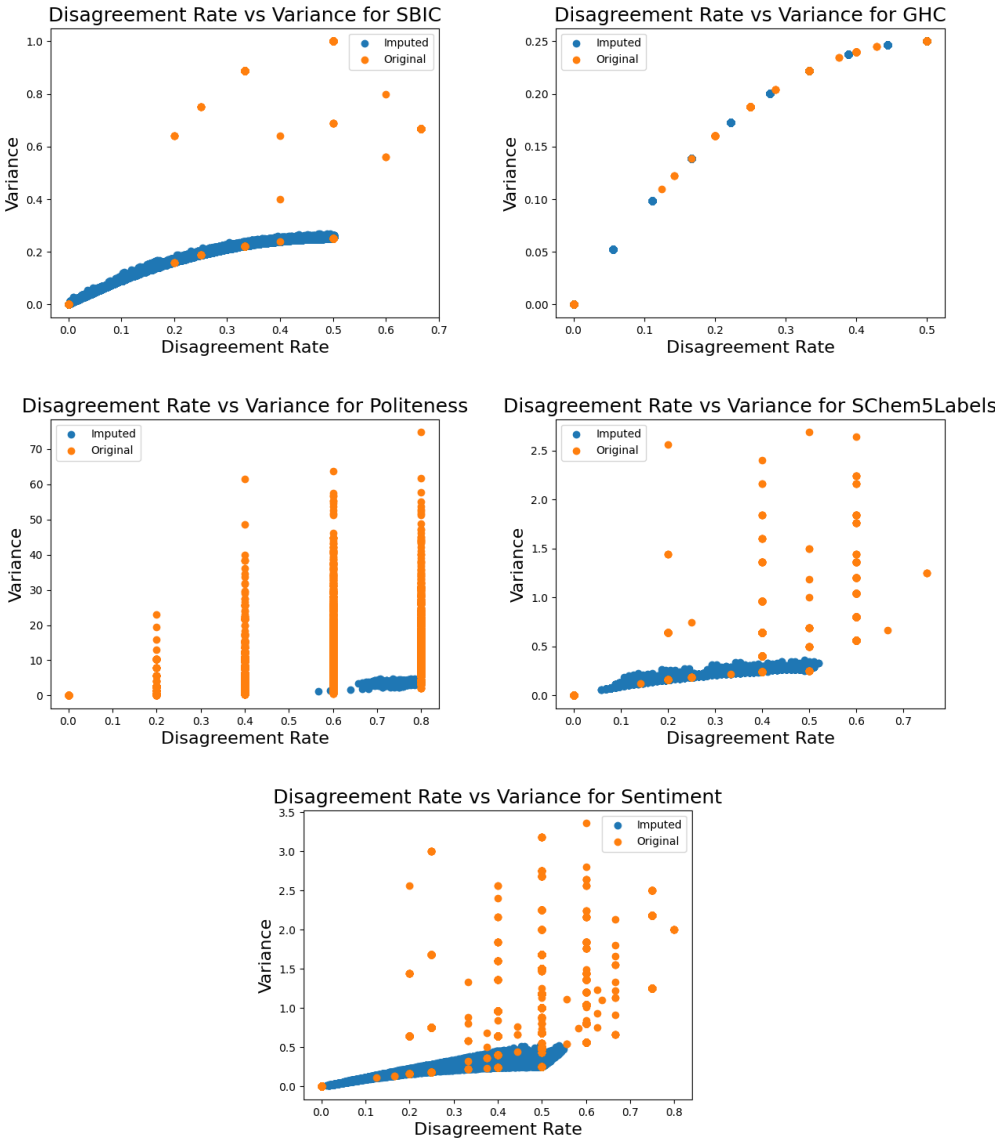

Figure 7: Visualizations of the decrease in variance after imputation. Orange is the original dataset, while blue is the imputed data. Each data point represents an example in the dataset. Variance is the variance among annotations for that example, and disagreement rate is the percentage of annotations that disagreed with the majority annotation. Vertical lines in the original dataset data appear because there are only a few annotators for most examples in the original datasets, meaning that the disagreement can only take on a few particular values.

Example 17192

['the whole me-being-republican thing must really bug these right wing nutjobs that dislike me so much.']

KL: **0.01 (Kernel)** vs 0.017 (NCF)

Original Distribution: [0. 0. 1.]

Kernel Distribution: [0. 0.01 0.99]

NCF Distribution: [0. 0.017 0.983]

Example 375

["Lord help me, I'm not trynna kill a cat for being annoying as fuck at 5 am"]

KL: 319 (Kernel) vs **1.415 (NCF)**

Original Distribution: [0.333 0. 0.667]

Kernel Distribution: [0.026 0.954 0.02]

NCF Distribution: [0.007 0.426 0.568]

Example 240

['people are so fucking ignorant stop trying to voice your opinion on a topic you know nothing about']

KL: **2.049 (Kernel)** vs 3.346 (NCF)

Original Distribution: [0.333 0. 0.667]

Kernel Distribution: [0.017 0.845 0.139]

NCF Distribution: [0.149 0.845 0.007]

Example 0

['what do you call a high person with downs syndrome ? a baked potato .']

KL: 0.1 (Kernel) vs **0.033 (NCF)**

Original Distribution: [0. 0.5 0.5]

Kernel Distribution: [0. 0.287 0.713]

NCF Distribution: [0.003 0.38 0.617]

Figure 8: Examples from the SBIC dataset.

Example 50

["I actually posted my comments a couple minutes ago. Do you think that explanation works, or do you think it's a good idea I elaborate further?"]

KL: 0.431 (Kernel) vs **0.321 (NCF)**

Original Distribution: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.2 0.2 0.4 0.2 0. 0. 0. 0. 0. 0. 0.]

Kernel Distribution: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.005 0. 0.005 0.041 0.11 0.228 0.311 0.164 0.078 0.032 0.018 0.005 0.005 0. 0. 0.]

NCF Distribution: [0. 0. 0. 0. 0. 0. 0. 0. 0.005 0. 0. 0. 0.014 0.032 0.078 0.169 0.288 0.237 0.082 0.05 0.032 0.009 0.005 0. 0.]

Example 49

['I see you have been creating lots of articles on bus stations on the <url>. Considering how little information there is (and is likely to be) for these stations, would it not be better to have them all on one article?']

KL: **0.55 (Kernel)** vs 0.641 (NCF)

Original Distribution: [0. 0. 0. 0. 0. 0. 0. 0. 0.2 0. 0. 0.4 0.2 0.2 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Kernel Distribution: [0. 0. 0. 0. 0. 0. 0. 0.005 0. 0.027 0.055 0.16 0.311 0.242 0.128 0.037 0.018 0.014 0. 0.005 0. 0. 0. 0. 0.]

NCF Distribution: [0. 0. 0. 0. 0. 0.005 0. 0.014 0.018 0.046 0.11 0.205 0.279 0.178 0.082 0.032 0.018 0.005 0.009 0. 0. 0. 0. 0. 0.]

Example 116

["Hi, I was just going to query your recent removal of the <url> article, without merging any of the article's content into <url>'s article. Was there any particular reason for the loss of content?"]

KL: **3.259 (Kernel)** vs 3.502 (NCF)

Original Distribution: [0. 0. 0. 0. 0. 0.2 0. 0. 0.2 0. 0. 0. 0. 0. 0. 0. 0. 0.2 0. 0.4 0. 0. 0. 0.]

Kernel Distribution: [0. 0. 0. 0. 0. 0.005 0. 0. 0.005 0.005 0. 0.027 0.073 0.174 0.342 0.201 0.096 0.023 0.027 0.009 0.014 0. 0. 0. 0.]

NCF Distribution: [0. 0. 0. 0. 0. 0.005 0.005 0. 0.005 0.009 0.014 0.068 0.114 0.219 0.297 0.119 0.082 0.032 0.018 0.005 0.009 0. 0. 0. 0.]

Example 60

["Thanks ! ...wouldn't you know my first award would be for being a smart ass ?"]

KL: 3.283 (Kernel) vs **2.566 (NCF)**

Original Distribution: [0. 0. 0. 0. 0. 0. 0. 0. 0.4 0. 0. 0. 0. 0. 0. 0. 0.4 0. 0.2 0. 0. 0. 0. 0.]

Kernel Distribution: [0. 0. 0. 0. 0. 0. 0. 0.005 0.009 0. 0.037 0.1 0.21 0.333 0.16 0.087 0.018 0.023 0.014 0.005 0. 0. 0. 0. 0.]

NCF Distribution: [0. 0. 0. 0. 0. 0.005 0. 0. 0.027 0.032 0.1 0.169 0.247 0.237 0.087 0.041 0.037 0.005 0.014 0. 0. 0. 0. 0. 0.]

Figure 9: Examples from the Politeness dataset.

Example 6207

["It's not okay to be to happy with a material object."]

KL: 3.016 (Kernel) vs **2.514 (NCF)**

Original Distribution: [0.6 0. 0. 0.4 0.]

Kernel Distribution: [0.029 0.235 0.716 0.02 0.]

NCF Distribution: [0.029 0.039 0.863 0.069 0.]

Example 58

["It's understandable if you use a customer's deposit to pay your shareholders' dividends."]

KL: 3.065 (Kernel) vs **2.964 (NCF)**

Original Distribution: [0.5 0.25 0. 0. 0.25]

Kernel Distribution: [0.02 0.02 0.892 0.059 0.01]

NCF Distribution: [0.02 0.029 0.814 0.127 0.01]

Example 10

["It's understandable if you're happy you have good genes."]

KL: **0.001 (Kernel)** vs 0.02 (NCF)

Original Distribution: [0. 0. 0. 0.8 0.2]

Kernel Distribution: [0. 0. 0. 0.814 0.186]

NCF Distribution: [0. 0. 0.02 0.784 0.196]

Example 3

["It's good to speak up for yourself."]

KL: **0.004 (Kernel)** vs 0.02 (NCF)

Original Distribution: [0. 0. 0. 0.8 0.2]

Kernel Distribution: [0. 0. 0. 0.833 0.167]

NCF Distribution: [0. 0. 0.02 0.775 0.206]

Figure 10: Examples from the SChem5Labels dataset.

Example 2

['!ubuntu Some how I now have both the old and new volume applets what package to I need to get rid of? screenshot: http://ur1.ca/5xeh']

KL: 0.073 (Kernel) vs **0.015 (NCF)**

Original Distribution: [0. 0.2 0.8 0. 0.]

Kernel Distribution: [0.001 0.093 0.885 0.022 0.]

NCF Distribution: [0. 0.217 0.769 0.014 0.]

Example 728

["@DCBadger Oh an afghan? Perhaps you're a sick old woman? ;) Feel better! Take vitamin C!"]

KL: **3.559 (Kernel)** vs 5.162 (NCF)

Original Distribution: [0. 0.5 0. 0. 0.5]

Kernel Distribution: [0. 0.025 0.808 0.159 0.008]

NCF Distribution: [0. 0.006 0.938 0.055 0.001]

Example 2642

["@avivhadar Oh, I am so so sorry for you and for the family! That's shocking and so young. Heartbreaking."]

KL: 2.033 (Kernel) vs **1.896 (NCF)**

Original Distribution: [0.25 0. 0.25 0.25 0.25]

Kernel Distribution: [0.002 0.053 0.81 0.13 0.005]

NCF Distribution: [0.001 0.128 0.668 0.192 0.011]

Example 2625

['@ashleej47 No. I can try selling some old prescription drugs for money.']

KL: 0.373 (Kernel) vs **0.0 (NCF)**

Original Distribution: [0. 0.75 0.25 0. 0.]

Kernel Distribution: [0.061 0.907 0.032 0. 0.]

NCF Distribution: [0. 0.753 0.247 0. 0.]

Figure 11: Examples from the Sentiment dataset.

Example 17054

['@Heartiste\xa0 ISIS is representative of Mossad. ']

KL: 0.365 (Kernel) vs **0.0 (NCF)**

Original Distribution: [0.667 0.333]

Kernel Distribution: [0.944 0.056]

NCF Distribution: [0.667 0.333]

Example 17052

['Lava Boils Away Largest Lake On Big Island In About 90 Minutes\xa0\xa0https://principia-scientific.org/lava-boils-away-largest-lake-on-big-island-in-about-90-minutes/']

KL: 0.0 (Kernel) vs 0.0 (NCF)

Original Distribution: [1. 0.]

Kernel Distribution: [1. 0.]

NCF Distribution: [1. 0.]

Example 9533

['Dudley sounds like an Islamized shit-hole.']

KL: 1.099 (Kernel) vs **0.251 (NCF)**

Original Distribution: [0. 1.]

Kernel Distribution: [0.667 0.333]

NCF Distribution: [0.222 0.778]

Figure 12: Examples from the GHC dataset.

Example 18

["It's expected that some people have trysts in airplanes."]

KL: 0.14 (Kernel) vs **0.037 (NCF)**

Original Distribution: [0.02 0.52 0.28 0.16 0.02]

Kernel Distribution: [0.01 0.34 0.54 0.1 0.01]

NCF Distribution: [0.01 0.48 0.39 0.11 0.01]

Example 3

["It's bad for a child to be in danger."]

KL: 0.001 (Kernel) vs **0.0 (NCF)**

Original Distribution: [0. 0. 0. 0.16 0.84]

Kernel Distribution: [0. 0. 0. 0.18 0.82]

NCF Distribution: [0. 0. 0. 0.17 0.83]

Example 2

['It is dangerous for a doctor to misdiagnose a patient.']

KL: 0.007 (Kernel) vs **0.002 (NCF)**

Original Distribution: [0. 0. 0. 0.22 0.78]

Kernel Distribution: [0. 0. 0. 0.27 0.73]

NCF Distribution: [0. 0. 0. 0.25 0.75]

Figure 13: Additional examples from the SChem dataset.